Experimental Evaluation of Sketching Techniques for Big Spatial Data

A. B. Siddique, Ahmed Eldawy
[msidd005,eldawy]@ucr.edu.
Department of Computer Science and Engineering, University of California, Riverside.

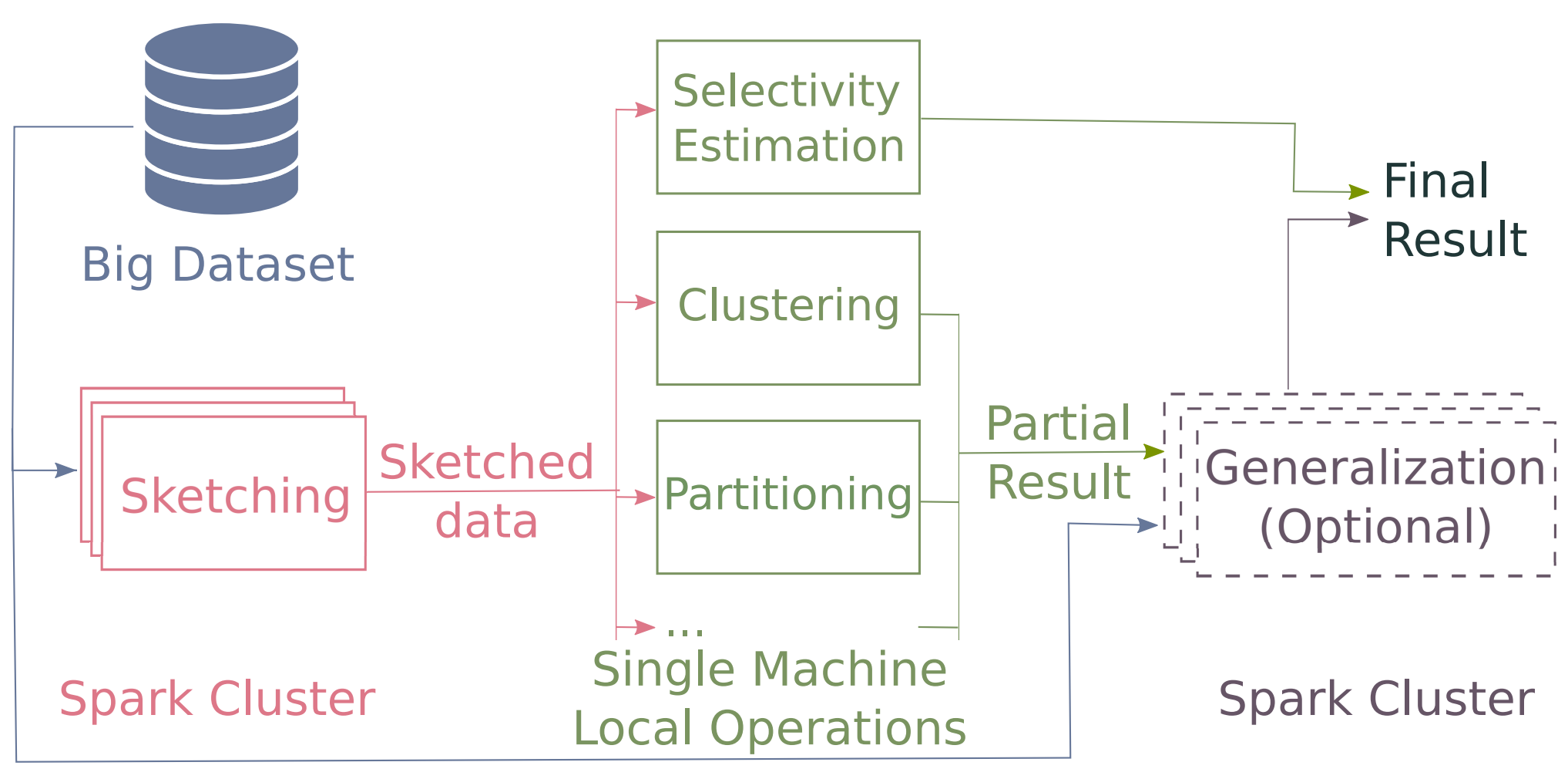
Motivation

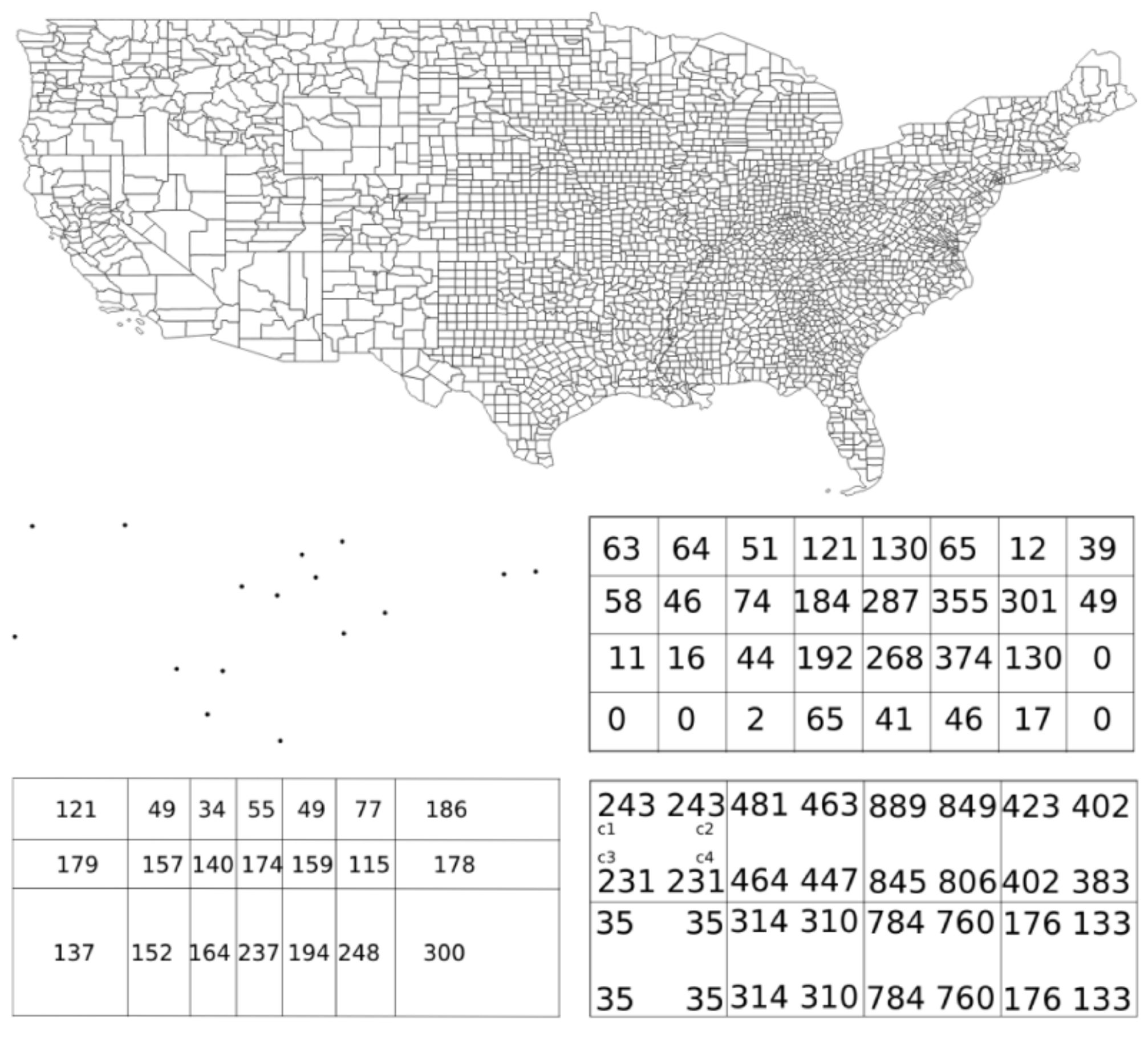
- Swift growth of the data
  - 2.5 exabytes of data is produced daily, of which 60 − 80% is geo-referenced.
  - Space telescopes broadcast about 140 GB data weekly.

- New scalable query processing techniques are need of the hour.
- Sketching techniques excluding sampling, are not well-studied due to two challenges.
  - Hard to compare their performance.
  - Might require some tweaks to the algorithms to work.
- A comprehensive evaluation to understand the trade-offs in the different sketching techniques for big spatial data.

Overview

- Three-phase sketching-based framework for big data processing.





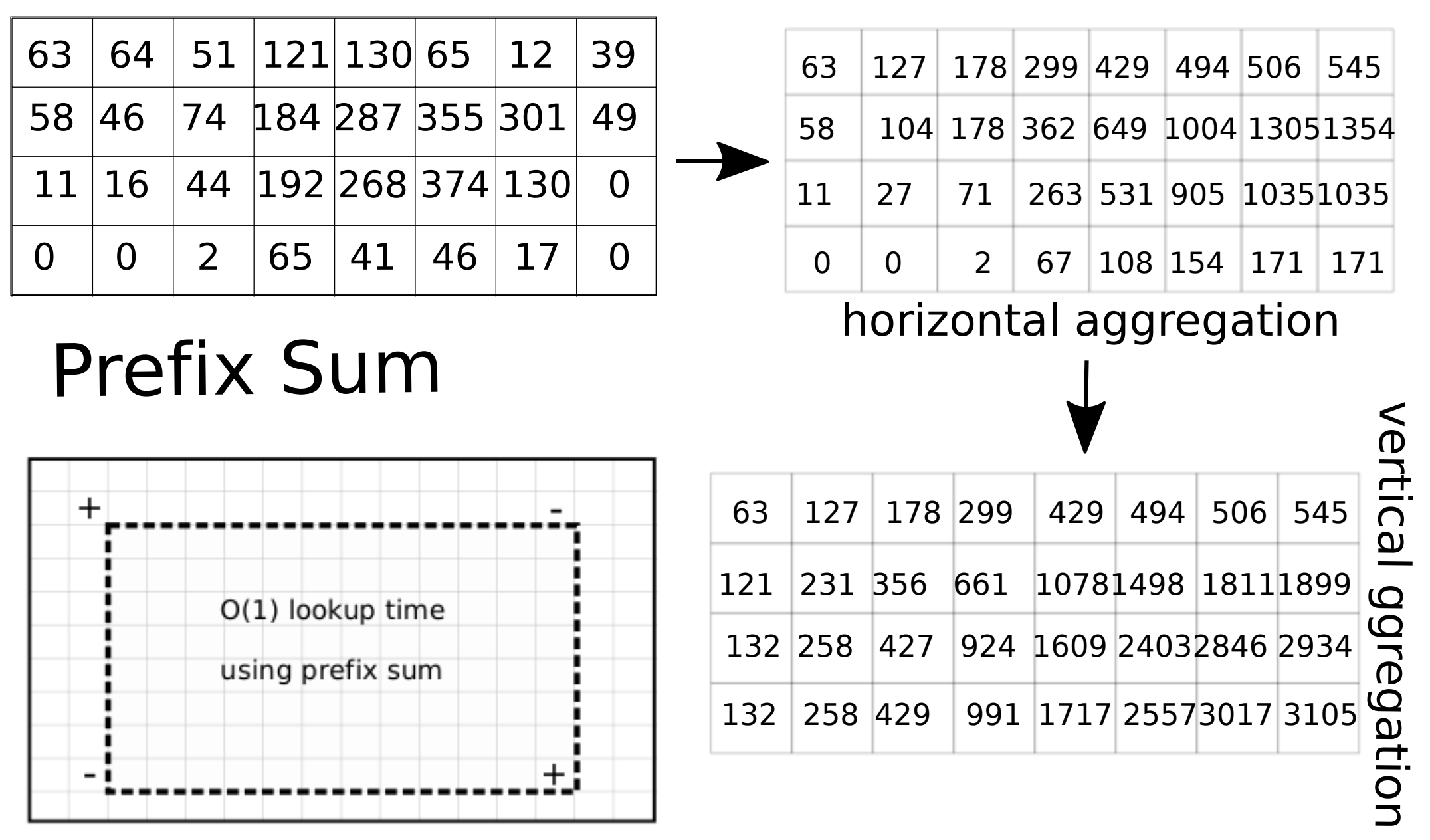
63	64	51	121	130	65	12	39
58	46	74	184	287	355	301	49
11	16	44	192	268	374	130	0
0	0	2	65	41	46	17	0

121	49	34	55	49	77	186
179	157	140	174	159	115	178
137	152	164	237	194	248	300

243 c1	243 c2	481	463	889	849	423	402
c3 231	c4 231	464	447	845	806	402	383
35	35	314	310	784	760	176	133
35	35	314	310	784	760	176	133

- Data is sketched only once for all future local operations.
- To make the sketching methods comparable, a parameter B is used.
- Local operations phase allows to reuse existing algorithm(s) with minimal changes.
- Optional generalization phase is merely a scan of the whole dataset in parallel.

Selectivity Estimation



63	64	51	121	130	65	12	39
58	46	74	184	287	355	301	49
11	16	44	192	268	374	130	0
0	0	2	65	41	46	17	0

63	127	178	299	429	494	506	545
58	104	178	362	649	1004	1305	1354
11	27	71	263	531	905	1035	1035
0	0	2	67	108	154	171	171

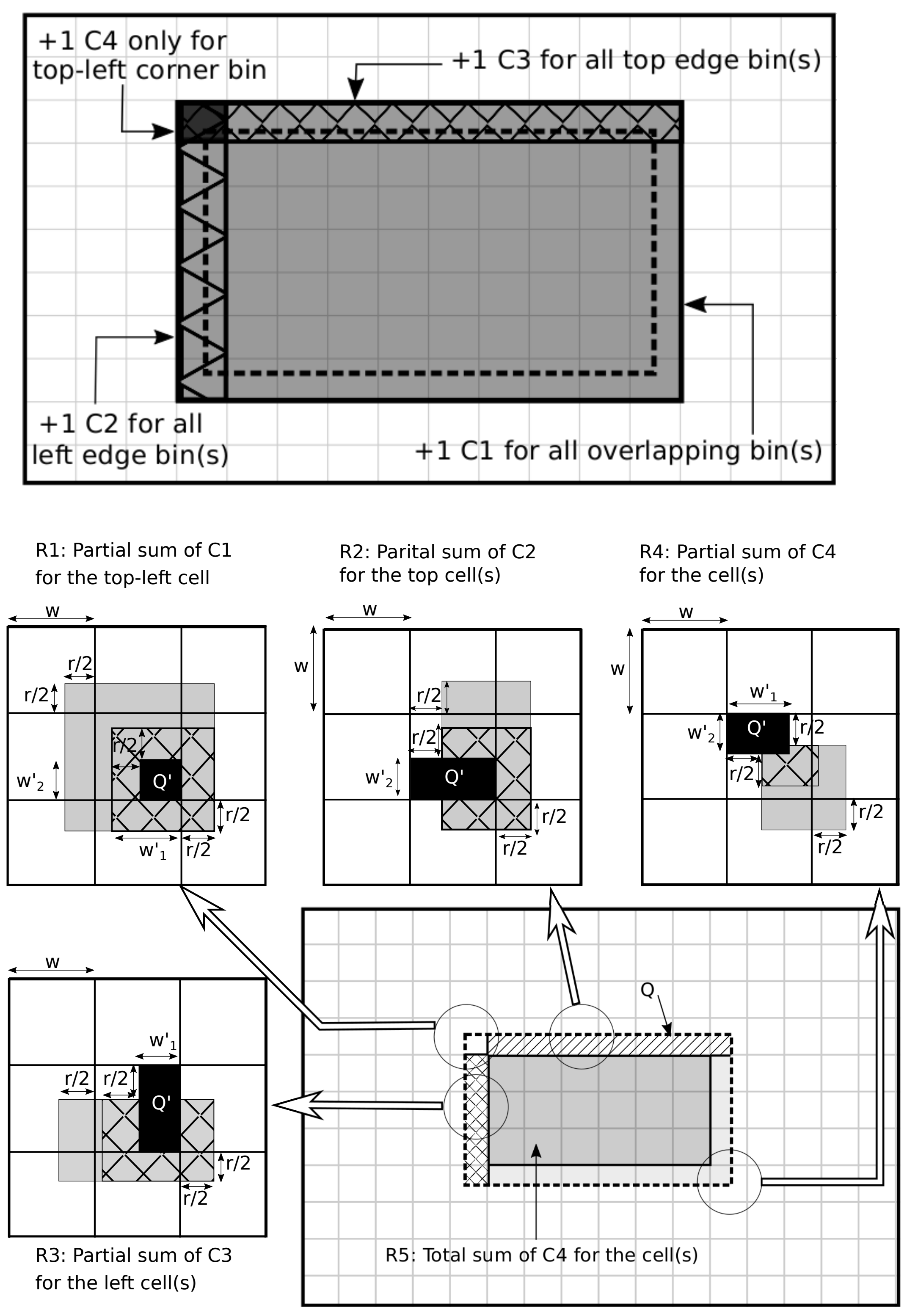
horizontal aggregation

63	127	178	299	429	494	506	545
121	231	356	661	1078	1498	1811	1899
132	258	427	924	1609	2403	2846	2934
132	258	429	991	1717	2557	3017	3105

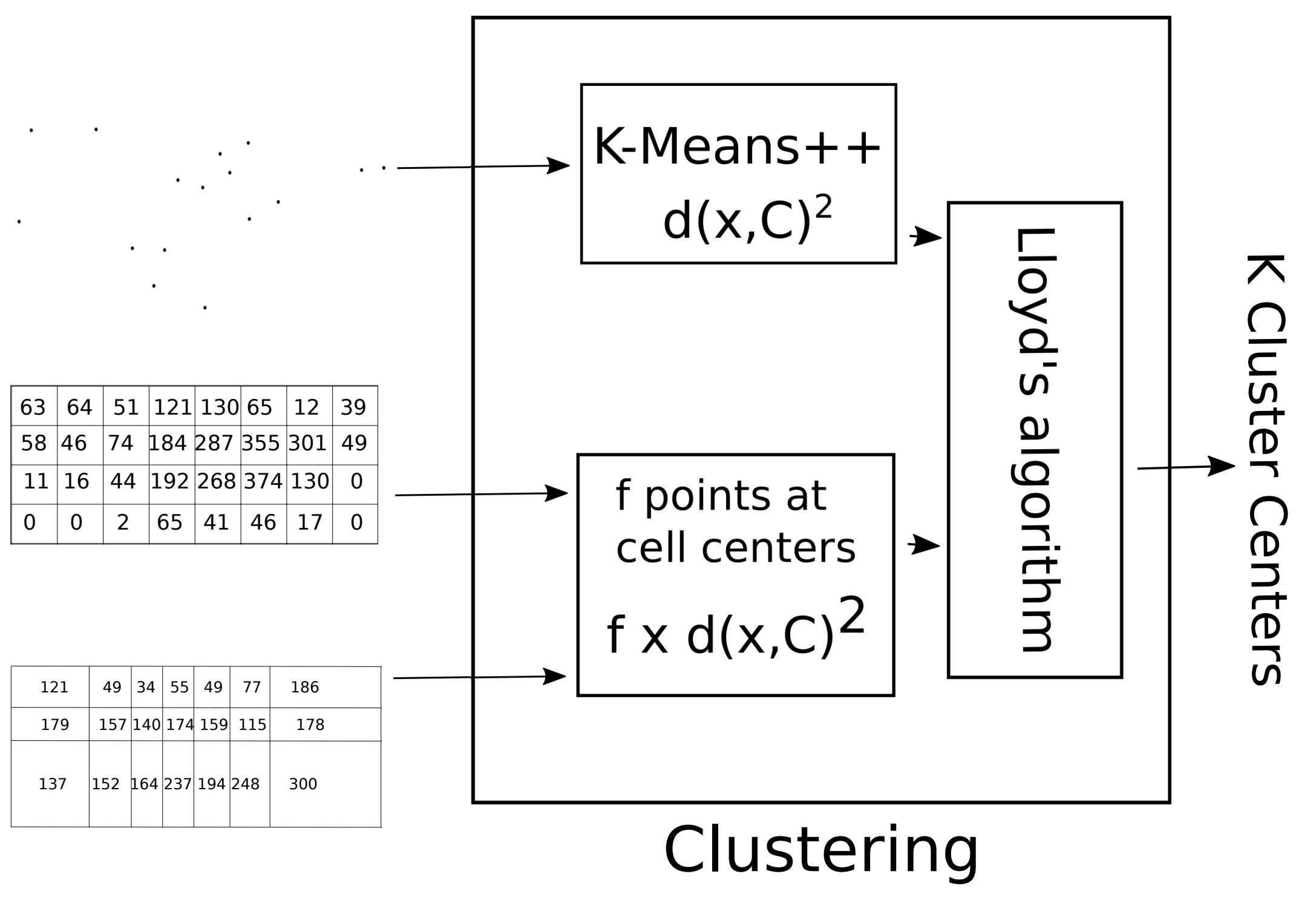
vertical ggregation

Prefix Sum

Euler Histogram



Clustering

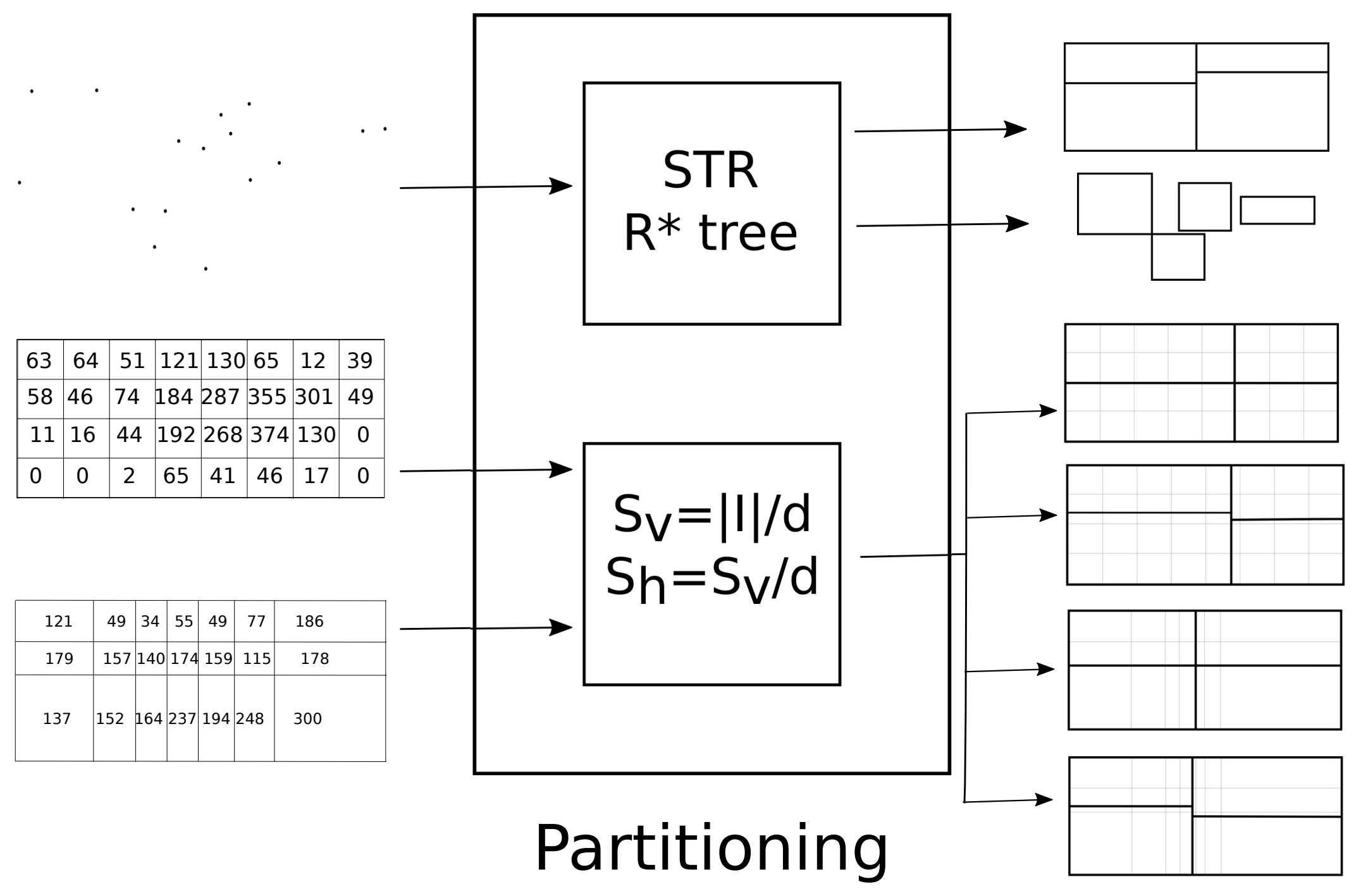


63	64	51	121	130	65	12	39
58	46	74	184	287	355	301	49
11	16	44	192	268	374	130	0
0	0	2	65	41	46	17	0

121	49	34	55	49	77	186
179	157	140	174	159	115	178
137	152	164	237	194	248	300

Clustering

Partitioning



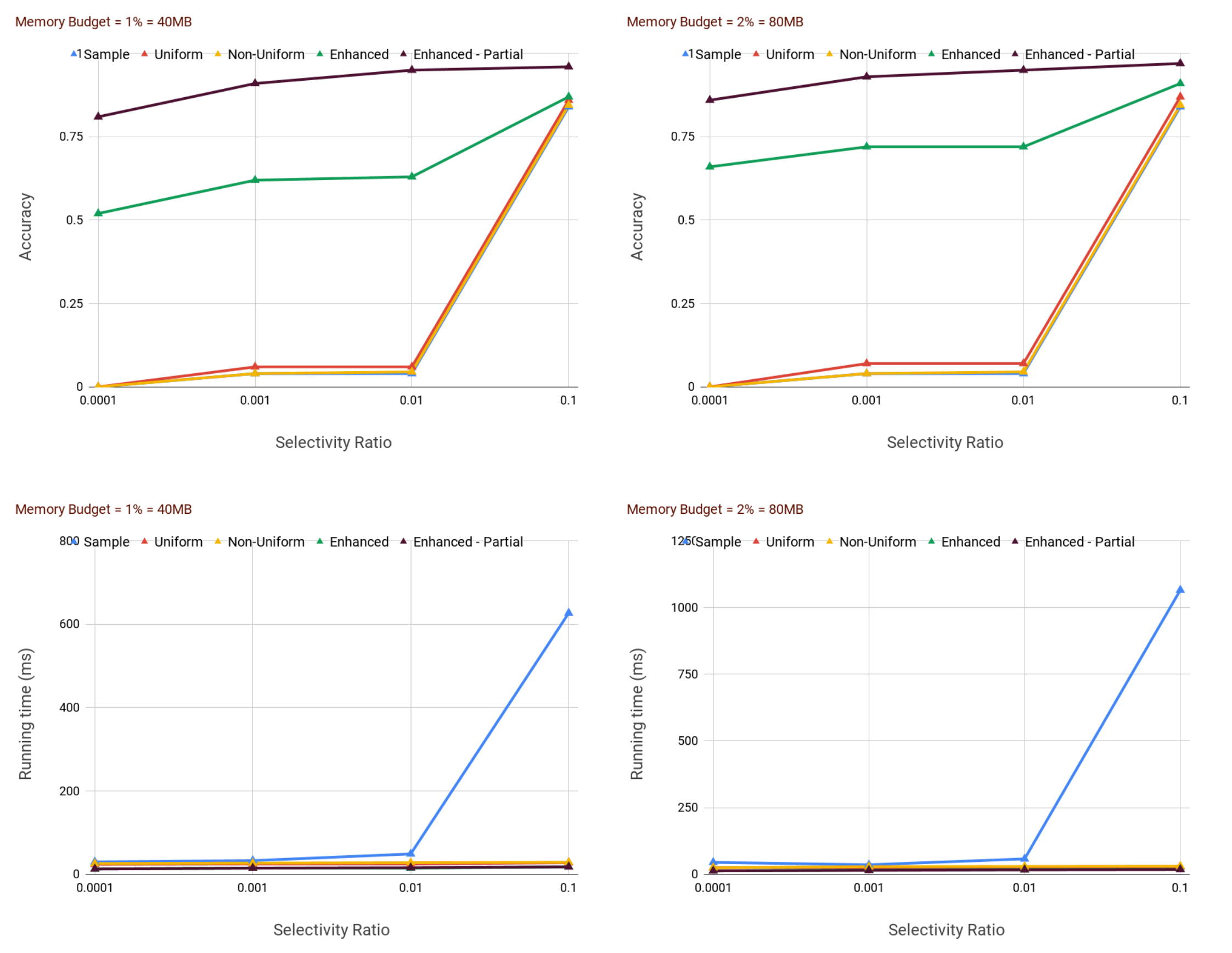
63	64	51	121	130	65	12	39
58	46	74	184	287	355	301	49
11	16	44	192	268	374	130	0
0	0	2	65	41	46	17	0

121	49	34	55	49	77	186
179	157	140	174	159	115	178
137	152	164	237	194	248	300

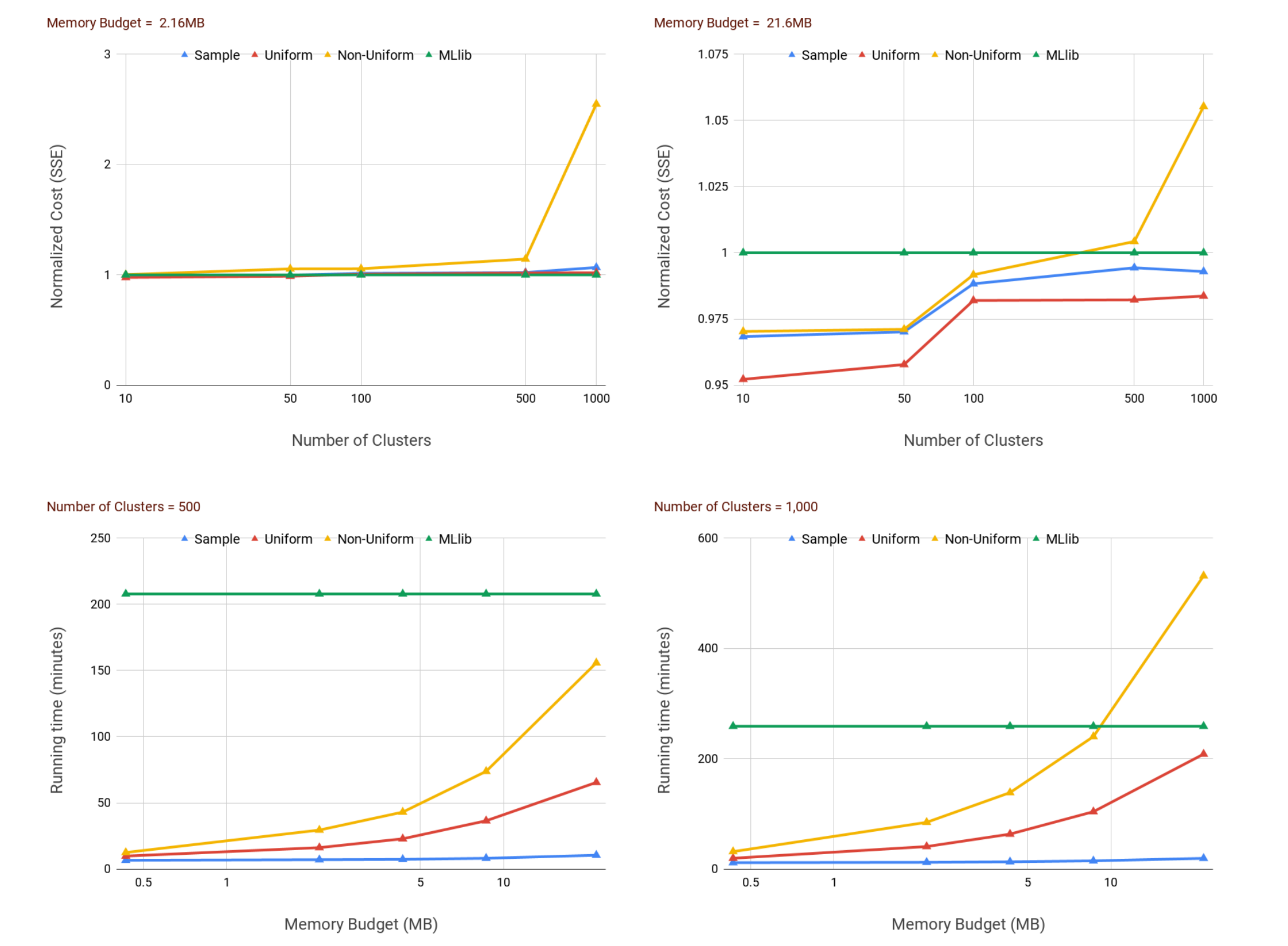
Partitioning

Experimental Evaluation

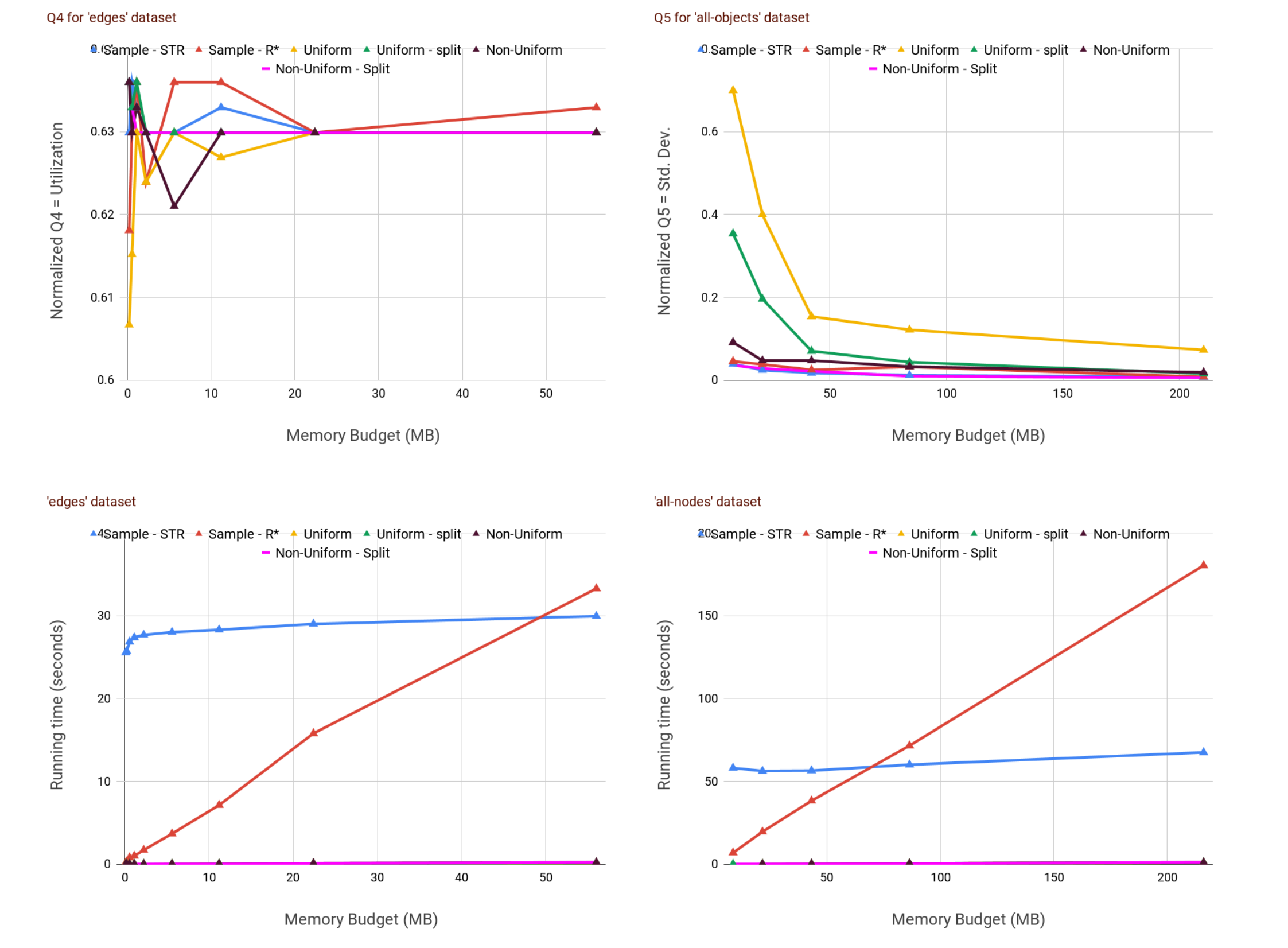
- **Selectivity Estimation**



- **Clustering**



- **Partitioning**



References

[1] Chasparis, Harry, and Ahmed Eldawy., "Experimental evaluation of selectivity estimation on big spatial data" in *Proceedings of the Fourth International ACM Workshop on Managing and Mining Enriched Geo-Spatial Data*, 2017, pp. 8. ACM.

[2] Bahmani, Bahman, et al., "Scalable k-means++" in *Proceedings of the VLDB Endowment*, 2012, pp. 622–633.

[3] Eldawy, Ahmed and Alarabi, Louai and Mokbel, Mohamed F, "Spatial partitioning techniques in SpatialHadoop" in *Proceedings of the VLDB Endowment*, 2015, pp. 1602–1605.

This work is supported in part by the
National Science Foundation (NSF)
under grant IIS-1838222.